Acta Scientiarum

http://periodicos.uem.br/ojs/acta
ISSN on-line: 1807-8656
Doi: 10.4025/actascihumansoc.v44i1.62246

FILOSOFIA

Zemblanity and Big Data: the ugly truths the algorithms remind us of

Ricardo Peraça Cavassane

Departamento de Filosofia, Centro de Lógica, Epistemologia e História da Ciência, Universidade Estadual de Campinas, Cidade Universitária Zeferino Vaz, s/n, 13083-970, Campinas, São Paulo, Brazil. E-mail: ricardo.peraca@gmail.com

ABSTRACT. In this paper, we will argue that, while Big Data enthusiasts imply that the analysis of massive data sets can produce serendipitous (that is, unexpected and fortunate) discoveries, the way those models are currently designed not only does not create serendipity so easily but also frequently generates zemblanitous (that is, expected and unfortunate) findings.

Keywords: serendipity; zemblanity; big data; data science; racism.

Zemblanidade e Big Data: as duras verdades das quais os algoritmos nos lembram

RESUMO. Neste artigo, argumentaremos que, enquanto os entusiastas dos Big Data sugerem que a análise de conjuntos de dados massivos pode produzir descobertas serendipitosas (isto é, inesperadas e benéficas), a maneira pela qual tais modelos são projetados atualmente não apenas não cria serendipidade tão facilmente, como também gera, frequentemente, resultados zemblanitosos (isto é, esperados e maléficos).

Palavras-chave: serendipidade; zemblanidade; big data; ciência de dados; racismo.

Received on January 25, 2022.
Accepted on April 7, 2022.

Introduction

The goal of this work is to argue that, while statistical models of social phenomena based on Big Data sometimes aim at finding genuine correlations that are surprising, valuable, and even both – and, thus, perhaps consisting in discoveries one could call serendipitous –, the results of those models can frequently be characterized as the diametrical opposite of serendipitous, that is, as zemblanitous: instead of allowing one to make fortunate and unexpected discoveries, those models often lead to unfortunate and expected results.

In section one, we will characterize zemblanity in detail from its initial definition, that is, as the polar opposite of serendipity, by the negation of the features found in detailed characterizations of serendipity. By doing this, we will be claiming that, in a certain research context, the chances of serendipitous discoveries and zemblanitous findings are inversely proportional.

In section two, we will argue that Big Data enthusiasts imply that the analysis of massive data sets can produce serendipitous discoveries, and will also argue that, the way statistical models of social phenomena based on Big Data are currently envisaged and designed, they do not generate serendipity as easily as their enthusiasts claim.

In section three, we will argue that statistical models of social phenomena based on Big Data, the way they are currently designed, frequently generate zemblanity, and will provide an example of zemblanitous findings in those models, namely the appearance of racist patterns.

Finally, in the closing remarks, we will claim that those models could be envisaged and designed in a way that would increase the chance of serendipitous discoveries and, consequently, would also decrease the occurrence of zemblanitous findings.

Zemblanity as the polar opposite of serendipity

> Serendipity. From *Serendip*, a former name of Ceylon, now Sri Lanka. A word coined by Horace Walpole, who had invented it based on a folktale, whose heroes were always making discoveries of things they were not in quest of. Ergo:

> serendipity, the faculty of making happy and unexpected discoveries by accident. So what is the opposite of Serendip, a southern land of spice and warmth, lush greenery and humming birds, sea-washed, sun-basted? Think of another world in the far north, barren, ice-bound, cold, a world of flint and stone. Call it *Zembla*. Ergo: zemblanity, the opposite of serendipity, the faculty of making unhappy, unlucky and expected discoveries by design. Serendipity and zemblanity: the twin poles of the axis around which we revolve (Boyd, 1998, p. 234-235, emphasis in original).

The term zemblanity was coined as the complementary antonym or polar opposite of serendipity: while the latter refers to a fortunate and unexpected discovery, the former refers to an unfortunate and expected finding. From the first definition of serendipity by Walpole (2011), the term has been refined (and the phenomenon investigated) by many scholars in order to reflect as accurately as possible its meaning in the context of scientific research, but its antonym has not received the same attention. In this section, we will enumerate a number of features attributed to serendipity and then, by the negation of each of those features, draw a detailed picture of zemblanity in the context of scientific research.

Serendipity is, first and foremost, a fortunate discovery, both in the sense of being achieved (to a certain extent) by chance and in the sense of being valuable. Concerning the second aspect (we will return to the first one later), a serendipitous discovery, in the context of scientific research, is valuable because it solves a problem or presents the means or opportunity to solve a problem, introduces a new and promising hypothesis or line of investigation, or even "[...] becomes the occasion for developing a new theory or for extending an existing theory" (Merton, 1948, p. 506). Zemblanity, on the other hand, is an unfortunate finding, though not in the sense of being brought by bad luck, for it does not involve chance at all, and also not contrary to valuable in the sense of being merely worthless, but in the sense of being undesirable. A zemblanitous finding reveals an underlying problem or issue, a negative side-effect or consequence, etc.

Secondly, serendipity is unexpected, not only in the sense of surprising but also in the sense of unpredictable, for it is unsought, that is, the researcher was not looking for it. Whether it solves a problem different than the one the researcher was concerned with (in the original sense of the word, which Yaqub (2018) calls Walpolian), or the exact problem the researcher was dedicated to, but in an unexpected way (in the sense he calls Mertonian, and which Roberts (1989) calls pseudoserendipity), or even if the researcher was not targeting a specific problem when it appears (in the senses Yaqub (2018) calls Bushian, if it solves an existing problem, or Stephanian, if it is seen as a potential solution to a problem to be determined later), a serendipitous discovery is always contingent. Zemblanity, by its turn, is either expected or predictable, depending on how informed the researcher is; that is, its appearance should not be a surprise to the researcher. This is why it is not a discovery at all, but rather a necessary finding one will eventually and inevitably arrive at when dealing with a certain problem or subject matter.

As we have stated before, a serendipitous discovery happens by chance or, to be more precise, "[...] at the intersection of chance and wisdom [...]" (Copeland, 2019, p. 2386), or made by accident and sagacity, as Walpole (2011) himself puts it. This means that, while it is triggered by a random event, only a skilled and attentive observer will be able to detect it, especially in true or Walpolian serendipity, in which the discovery concerns a problem extrinsic to the subject matter targeted by the research. A zemblanitous finding is in no way the product of chance, for it happens, as Boyd (1998) states, by design; in other words, it is a feature, not a bug, and it concerns a problem intrinsic or closely related to the subject matter targeted by the research. Consequently, only the unskilled and negligent will not be prepared for its appearance.

The characteristics of serendipity and zemblanity we have analyzed thus far are already implied in their original definitions; the ones we will analyze from now on are the result of the investigations of scholars on the phenomenon of serendipity, especially in the context of the sciences, and of our further characterization of zemblanity by opposition to serendipity.

As we have seen, a serendipitous discovery is unpredictable, not only to the researcher who made the discovery but also to an observer who tried to predict the outcomes of the interaction between a certain researcher and a certain problem in a certain context; that is, serendipity is an emergent phenomenon. It is made possible by a conjunction of many different factors – including methodological, behavioral, and environmental ones – in a complex dynamic with the presence of self-organization and circular causality whose product is unforeseeable, for it is not reducible to the sum of those interactions. Thus, as Copeland (2019) puts it, serendipity is not merely the result of weak emergence, that is, not merely unexpected because of our epistemic limitations, but of strong emergence, that is, unpredictable due to the complexity of the context in which it appears. Consequently, serendipity is only retrospectively recognizable, "[...] once the valuable outcome has been determined and upon reflection on the now-apparent significance of the relevant unexpected events and insight" (Copeland, 2019, p. 2391-2392). A zemblanitous finding, contrarily, is the result of linear causality, of a mechanical interaction between factors in a simple structure with a strong presence of hetero-organization that is predictable and, thus, can be prospectively identified.

Yaqub (2018, p. 172) divides those factors into four categories: "Serendipity may be theory-led, observer-led, error-borne or network-emergent". All of those factors depend on certain characteristics of the researcher, the researcher's environment, and the interaction between both to be able to generate serendipitous discoveries. On the individual researcher's side, according to Merton (1984), it takes knowledge and skill to identify an unanticipated datum as anomalous, that is, as "[...] inconsistent with prevailing theory or with other established facts [...]" (Merton, 1984, p. 506) and as strategic, that is, "[...] that it must permit of implications which bear upon generalized theory" (Merton, 1984, p. 507). Thus, serendipity requires attentiveness to detail, openness to error, preparedness to deal with anomalous by-products, and inquisitiveness to pursue their possible implications. Yaqub (2018) also notes that "observation routines and instrumentation ... may affect an observer's perceptiveness" (Yaqub, 2018, p. 173). Such characteristics are not only pertinent to individual researchers, but also to their environment, which can provide the researcher with a network that may bring "[...] discoveries to the attention of researchers who can exploit them [...]" (Yaqub, 2018, p. 174) and "[...] where exploitation of an observation may require the skills and resources of multiple people" (Yaqub, 2018, p. 174). Thus, the diversity and flexibility of the researcher's environment increase the chance of serendipitous discoveries: "In communities that allow diverse members opportunities to contribute to the production of knowledge, there will both be more opportunities for serendipitous discovery and community members will develop the skills necessary to take advantage of those opportunities" (Copeland, 2019, p. 2386).

On the other hand, findings that consist of normal, consistent data, are only considered zemblanitous when not anticipated, and thus are the result of a researcher's ignorance, negligence, unpreparedness, or excessive self-confidence and of a restrictive and rigid environment, which presents the researcher with a single, isolated object or problem. Yaqub (2018) also notices that "[...] networks that are particularly homogeneous, cohesive and insular may be inversely related to serendipity [...]" (Yaqub, 2018, p. 174), from which we may deduce they could promote zemblanity.

The search for serendipity in Big Data

When it analyzes structured datasets and small numbers of variables, data science relies on basic statistics to deal with simple problems. However, when the datasets consist of massive, rapidly growing volumes of semi-structured or unstructured data of various different types, the methods and goals of the so-called Big Data analytics also change. Mayer-Schönberger and Cukier (2013) define Big Data as "[...] the ability ... to harness information in novel ways to produce useful insights or goods and services of significant value [...]" (Mayer-Schönberger & Cukier, 2013, p. 2) and explain why it is, in their view, revolutionary:

> Big Data is about what, not why. We don't always need to know the cause of a phenomenon; rather, we can let data speak for itself. Before Big Data, our analysis was usually limited to testing a small number of hypotheses that we defined well before we even collected the data. When we let the data speak, we can make connections that we had never thought existed (Mayer-Schönberger & Cukier, 2013, p. 14).

This is also the opinion of Anderson (2008): "The new availability of huge amounts of data, along with the statistical tools to crunch these numbers, offers a whole new way of understanding the world. Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all".

If Big Data worked the way its enthusiasts claim, it would be a perfect 'serendipity machine': the algorithmic analysis of a data set so massive and diverse it covers almost every instance and aspect of a certain phenomenon – what Mayer-Schönberger and Cukier (2013) call 'N=all' –, would produce a great number of correlations – many of them spurious, but some of them genuine and therefore potentially valuable – some of which the researcher would never expect to find, a product of the algorithm's capacity to uncover hidden patterns and of the richness of relations between the different types of data, gathered and analyzed without human interference and bias. Thus, the Big Data analyst, whether focused on a specific problem or not, would probably be able to make one or even many serendipitous discoveries, which could also depend on the researcher's ability to detect the most valuable correlations.

However, Big Data usually does not work the way its enthusiasts claim, mainly because it fails to represent a phenomenon in its totality and from a neutral, theoretically independent perspective. According to Schutt and O'Neil (2014), it is practically never possible to reach 'N=all' and frequently Big Data lets pass precisely the most relevant cases. An example of a source of data widely used by those who intend on identifying social patterns through Big Data is twitter. As Boyd and Crawford (2012) notice:

> Twitter does not represent 'all people', and it is an error to assume 'people' and 'Twitter users' are synonymous: they are a very particular sub-set. Neither is the population using Twitter representative of the global population. Nor can we assume that accounts and users are equivalent. Some users have multiple accounts, while some accounts are used by multiple people. [...] Some accounts are 'bots' that produce automated content without directly involving a person (Crawford, 2012, p. 669, emphasis in original).

Taking the totality of the tweets on a certain subject (selected through certain keywords) as a proxy for the totality of public manifestations on the subject is a huge mistake. The impossibility of representing the totality of occurrences in a Big Data-based model, however, is not restricted to social networks. According to Kitchin (2014), although such a model may intend to be exhaustive and neutral, it is "[...] both a representation and a sample, shaped by the technology and platform used, the data ontology employed and the regulatory environment, and it is subject to sampling bias" (Kitchin, 2014, p. 4). Thus, the suggestion that the unfocused analysis of massive datasets can replace scientific research guided by hypotheses and grounded in theory is epistemologically problematic – as Ibekwe-SanJuan and Bowker (2017, p. 194) highlight, the proprietary nature of the data and the opacity of the algorithms go against "[...] the principles of falsifiability and fallibilism ... which have guided scientific activity up till now [...]" – and also deceiving – for Big Data-based models incorporate many presuppositions, whether when collecting, selecting, analyzing or visualizing data.

Behind the belief that Big Data would allow us to abandon any theories is the mistake that the datum represents, ipsis litteris, that which is given in nature. In reality, there is no 'raw data', or, as Bowker (2005, p. 184, emphasis in original) puts it: "'Raw data' is both an oxymoron and a bad idea". That is, even if Big Data could somehow represent the totality of the occurrences of a phenomenon, the data would not "[...] speak for themselves [...]", unless not in the sense that data could be neutral or free from theoretical presuppositions. Furthermore, Big Data-based models would not eliminate the biases present in human-made analyses, as Crawford (2013) highlights:

> Can numbers actually speak for themselves? Sadly, they can't. Data and data sets are not objective; they are creations of human design. We give numbers their voice, draw inferences from them, and define their meaning through our interpretations. Hidden biases in both the collection and analysis stages present considerable risks, and are as important to the big-data equation as the numbers themselves.

Thus, apart from those contexts in which the datafication of the totality of the instances is really possible, causal explanations are unnecessary and errors are tolerable – for instance when a streaming service analyzes every customer interaction in its platform to suggest a movie for a specific customer –, in which Big Data aims at solving problems that Auerbach (2014) calls problems of 'selection optimization', Big Data can be successful and surprising discoveries are possible, even though their value and, therefore, serendipitous character, is debatable. However, when it deals with complex and sensitive issues in which none of the criteria listed above can be met, and in which unexpected discoveries can be really valuable, Big Data often fails. Thus, models built under this view of how Big Data works not only do not generate serendipity so easily but also, as we will discuss, frequently create zemblanity.

The finding of zemblanity in Big Data

Social phenomena are complex. In the words of Crawford, Miltner, and Gray (2014, p. 1667), "Aggregated, individual actions cannot, in and of themselves, illustrate the complicated dynamics that produce social interaction – the whole of society is greater than the sum of its parts". Such phenomena and their complexity have been the subject of the social sciences; however, the enthusiasts of Big Data claim that, provided with massive enough datasets and powerful algorithms, the statistical analysis of those data alone can produce more knowledge on social phenomena than the traditional social sciences do. As we have argued, Big Data-based models are frequently unsuccessful when they fail to represent a phenomenon in its totality of relevant instances and aspects, and when they deal with complex and error-sensitive issues. It escapes from the scope of this work to elaborate on how data science is inherently incapable of surpassing traditional science in the understanding of complex systems, so here we will focus on showing, through some examples, how Big Data-based models fail to properly represent some social phenomena, while at the same time exhibiting an aura of success and scientificity. While doing so, we will argue that the way those models are envisaged and designed leads to those failures and produces zemblanity.

Many of those examples come from the areas of public and private security, specifically from surveillance, policing, and incarceration. Researchers, including Brayne (2021), Ferguson (2017), and O'Neil (2016), have shown the flaws and biases (especially racial ones) present in models of face recognition, predictive policing, and recidivism risk assessment. PredPol, for instance, one of the many predictive policing models currently in use in the United States, processes historical crime data and calculates, hour by hour, where crimes are most likely to occur; its algorithm "[...] looks at a crime in one area, incorporates it into historical patterns, and predicts when and where it might occur next" (O'Neil, 2016, p. 88). Not focusing on individuals, PredPol would supposedly not be influenced by racial biases.

However, "To understand the development of predictive policing, it must be situated within the policing reform movement in the United States aimed at making the police more proactive and vigilant, rather than reactive and emergency-focused" (Benbouzid, 2019, p. 1). Directed by zero-tolerance policies, stop-and-frisk practices, and productivity quotas, police departments feed PredPol not only with violent crime data but also and mostly with data on minor crimes, like illicit drug use. A study (Lum & Isaac, 2016) has shown that the data on drug use produced by the police is by far not representative of the reality of drug use, being reported mainly in impoverished neighborhoods, where the majority of the population is African-American or Hispanic. When that data is fed to the algorithm of PredPol, it directs the police to those neighborhoods, concentrating crime-reporting on those areas, generating a feedback loop, which not only reinforces the biases already present in the data but also corroborates the biased results, granting them an aura of scientificity. Fed by data itself helps create, PredPol behaves as a self-fulfilling prophecy:

> PredPol ... wins at every turn! PredPol will announce that a crime is to take place in a specific area of the city. Off the policeman goes to respond to the situation. One of two things will happen: either a crime takes place as planned and the policeman stops the offender, in which case the PredPol software receives its gold star; or no offence occurs. But this is probably linked to the on the spot presence of the policeman, and so it is still a gold star for the software. We cannot blame PredPol, which prevented the crime (Dupuy, 2018, p. 160).

For Benslimane (2014), Predpol ignores the various sociological factors that lead to criminality and the biases in policing to create a simplified, apparently objective representation according to which there are more crimes in certain areas of a city. As we have argued, it fails to accurately represent the actual presence of criminality in a city because the data it is fed with does not represent the totality of crimes that occur in that city (or even a random sample, but an arbitrary one), nor is the model informed by theories about why certain crimes are more common in certain areas than in others. Additionally, it is subject to a pernicious feedback loop.

Another example is the COMPAS model, designed to assess potential recidivism risk, which is being used in the United States to inform judges' decisions. A study (Larson, Mattu, Kirchner, & Angwin, 2016) has shown that in that model, whose rate of success is of merely approximately sixty percent, African-American defendants are frequently wrongly classified with a higher risk of recidivism, while Caucasian defendants are frequently wrongly classified with a lower risk of recidivism. A subsequent study (Flores, Bechtel, & Lowenkamp, 2016) argued that the study by Larson et al. (2016) did not accurately represent the reality of the COMPAS results; however, even though the differences between the failure rates are smaller than Larson et al. (2016) claim they are, they are still there. Flores et al. (2016) also fail to account for the racial biases in policing and for the feedback loop that a higher rate of incarceration of African-Americans creates on the rate of recidivism.

Also fed with biased data produced by the police and by the judicial system, the COMPAS model does not represent the totality of the instances of the phenomena it aims to model. Differently from the PredPol model, it is allegedly grounded in theory – various criminal theories are cited in the official guide to the use of the platform (Northpointe, 2015, p. 5-6); however, in the questionnaire that feeds the system with data from the defendants there are questions like 'How many of your friends/acquaintances are taking drugs illegally?' and 'How often did you get in fights while at school?'. The questionnaire also asks people to agree or disagree with statements such as 'A hungry person has a right to steal' (Angwin, Larson, Mattu, & Kirchner, 2016), among other questions of a highly subjective character. Thus, the model does not seem to be grounded in scientific theories that take sociological factors into account, nor is it concerned with what causes recidivism.

Thus, both models fail to accurately represent, respectively, the reality of geographical crime distribution and of potential recidivism, but their failure can be represented as success since their use increases certain indicators, their errors are not easily identifiable, they can behave as self-fulfilling prophecies, and especially since they reproduce prejudices and misconceptions already present in our society and seen as factual.

Models like these, instead of generating serendipitous discoveries, that is, instead of showing unexpected correlations that can have valuable applications – in this case, that would help prevent violent crimes –, generate zemblanitous findings: their results present racially biased patterns that an individual informed by the relevant theories would rightfully expect to encounter, but which nonetheless are not accounted for by the model. Their results do not deviate much from the already established practices, and thus their value lies not in the new insights they generate, but in how they corroborate and justify those practices to the public. However, their employment is not innocuous: the algorithms strengthen and naturalize the existent biases.

Final considerations

As we have seen, in the context of scientific practice, serendipity can be cultivated and, consequently, zemblanity can be avoided, and that could also be the case in the context of Big Data analysis. If Big Data-based models were fed with reliable, comprehensive data, were consistent with established scientific theories, were transparent and open to scrutiny, and accounted for the complexities of the phenomena analyzed, while including among the analysts independent researchers from diverse backgrounds, who could pursue open-ended goals in a flexible research environment, then serendipitous discoveries would be more likely to occur as the result of those models and, consequently, zemblanitous findings would be less common. It is when Big Data is envisaged as a cheaper, faster, and better substitute for traditional science, in complete disconnection with the best scientific practices, that it becomes a mere 'zemblanity machine'.

References

Anderson, C. (2008, June 23). The end of theory: will the data deluge make the scientific method obsolete? *Wired Magazine,* Science. Retrieved from https://www.wired.com/2008/06/pb-theory

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). Machine bias: there's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica*. Retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Auerbach, D. (2014, August 7). The big data paradox: it's never complete, and it's always messy – and if it's not, you can't trust it. *Slate*, Technology. Retrieved from http://www.slate.com/articles/technology/bitwise/2014/08/what_is_big_data_good_for_incremental_change_not_big_para digm_shifts.html

Benbouzid, B. (2019). To predict and to manage. Predictive policing in the United States. *Big Data & Society, 6*(1), 1-13. DOI: https://doi.org/10.1177/2053951719861703

Benslimane, I. (2014, décember 10). Étude critique d'un système d'analyse prédictive appliqué à la criminalité: Predpol. *CorteX Journal*. Retrieved from https://cortecs.org/politique-societe/predpol-predire-des-crimes-ou-des-banalites

Bowker, G. C. (2005). *Memory practices in the sciences*. Cambridge, MA: MIT Press.

Boyd, D., & Crawford, K. (2012). Critical questions for big data. *Information, Communication & Society, 15*(5), 662-679. DOI: https://doi.org/10.1080/1369118X.2012.678878

Boyd, W. (1998). *Armadillo*. London, UK: Hamish Hamilton.

Brayne, S. (2021). *Predict and surveil: data, discretion, and the future of policing*. New York, NY: Oxford University Press.

Copeland, S. (2019). On serendipity in science: discovery at the intersection of chance and wisdom. *Synthese, 196*(6), 2385-2406. DOI: https://doi.org/10.1007/s11229-017-1544-3

Crawford, K. (2013, April 1). The hidden biases in Big Data. *Harvard Business Review,* Analytics and data science. Retrieved from http://blogs.hbr.org/cs/2013/04/the_hidden_biases_in_big_data.html

Crawford, K., Miltner, K., & Gray, M. L. (2014). Critiquing big data: politics, ethics, epistemology. *International Journal of Communication, 8*(1), 1663-1672.

Dupuy, J-P. (2018).Science without philosophy: the case of big data. *Crisis and Critique, 5*(1), 147-161.

Ferguson, A. G. (2017). *The rise of big data policing: surveillance, race, and the future of law enforcement*. New York, NY: New York University Press.

Flores, A. W., Bechtel K., & Lowenkamp C. T. (2016). False Positives, false negatives, and false analyses: a rejoinder to 'machine bias: there's software used across the country to predict future criminals. And It's Biased Against Blacks'. *Federal Probation Journal, 80*(2), 38-46.

Ibekwe-Sanjuan, F., & Bowker, G. C. (2017). Implications of big data for knowledge organization. *Knowledge Organization, 44*(3), 187-198.

Kitchin, R. (2014). Big data, new epistemologies and paradigm shifts. *Big Data & Society, 1*(1), 1-12. DOI: https://doi.org/10.1177/2053951714528481

Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016, May 23). How we analyzed the COMPAS recidivism algorithm. *ProPublica*. Retrieved from https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Lum, K., & Isaac, W. (2016). To predict and serve? *Significance, 13*(5), 14-19. DOI: https://doi.org/10.1111/j.1740-9713.2016.00960.x

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: a revolution that will transform how we live, work and think*. New York, NY: Houghton Mifflin Harcourt.

Merton, R. K. (1948). The bearing of empirical research upon the development of social theory. *American Sociological Review, 13*(5), 505-515. DOI: https://doi.org/10.2307/2087142

Northpointe. (2015). *Practitioner's guide to COMPAS core*. Retrieved from https://assets.documentcloud.org/documents/2840784/Practitioner-s-Guide-to-COMPAS-Core.pdf

O'Neil C. (2016).*Weapons of math destruction: how big data increases inequality and threatens democracy*. New York, NY: Crown Publishers.

Roberts, R. M. (1989). *Serendipity: accidental discoveries in science*. New York, NY: John Wiley & Sons.

Schutt, R., & O'Neil C. (2014). *Doing data science*. Sebastopol, CA: O'Reilly Media.

Walpole, H. (2011). *Horace Walpole's correspondence: Yale edition*. Retrieved from https://libsvcs-1.its.yale.edu/hwcorrespondence/

Yaqub, O. (2018). Serendipity: towards a taxonomy and a theory. *Research Policy, 47*(1), 169-179. DOI: https://doi.org/10.1016/j.respol.2017.10.007